Submitted to the Annals of Applied Probability

SMALL WORLD MCMC WITH TEMPERING: ERGODICITY AND SPECTRAL GAP

BY YONGTAO GUAN* AND MATTHEW STEPHENS

Baylor College of Medicine and University of Chicago

When sampling a multi-modal distribution $\pi(x)$, $x \in \mathbb{R}^d$, a Markov chain with local proposals is often slowly mixing; while a Small-World sampler (Guan and Krone, 2007) – a Markov chain that uses a mixture of local and long-range proposals – is fast mixing. However, a Small-World sampler suffers from the curse of dimensionality because its spectral gap depends on the volume of each mode. We present a new sampler that combines tempering, Small-World sampling, and producing long-range proposals from samples in companion chains (e.g. Equi-Energy sampler). In its simplest form the sampler employs two Small-World chains: an exploring chain and a sampling chain. The exploring chain samples $\pi_t(x) \propto \pi(x)^{1/t}$, $t \in [1, \infty)$, and builds up an empirical distribution. Using this empirical distribution as its long-range proposal, the sampling chain is designed to have a stationary distribution $\pi(x)$. We prove ergodicity of the algorithm and study its convergence rate. We show that the spectral gap of the exploring chain is enlarged by a factor of t^d and that of the sampling chain is shrunk by a factor of t^{-d}. Importantly, the spectral gap of the exploring chain depends on the "size" of $\pi_t(x)$ while that of sampling chain does not. Overall, the sampler enlarges a severe bottleneck at the cost of shrinking a mild one, hence achieves faster mixing. The penalty on the spectral gap of the sampling chain can be significantly alleviated when extending the algorithm to multiple chains whose temperatures $\{t_k\}$ follow a geometric progression. If we allow $t_k \to 0$, the sampler becomes a global optimizer.

1. Introduction. Developing an algorithm to improve sampling efficiency for a high dimensional multi-modal distribution has been an active research area (Geyer, 1991; Marinari and Parisi, 1992; Neal, 2001; Kou *et al.*, 2006; Madras and Zheng, 2003; Guan and Krone, 2007; Andrieu *et al.*, 2008; Woodard *et al.*, 2008; Brockwell *et al.*, 2010; Del Moral and Doucet, 2010). In this paper we introduce a new sampling algorithm and study its ergodicity and convergence properties. The new algorithm combines several existing ideas: tempering (Geyer, 1991; Marinari and Parisi, 1992), propagating information between chains via empirical distributions (Kou *et al.*, 2006), and Small-World sampling (Guan *et al.*, 2006; Guan and Krone, 2007). A Small-World sampling combines local and long-range proposals in a Metropolis-Hastings algorithm. Guan and Krone (2007) showed

*Thanks to Winfried Barta for helpful discussions regarding ergodicity proof.

AMS 2000 subject classifications: Primary 65C05, ; secondary 65C40

Keywords and phrases: Markov Chain, Monte Carlo, Small World Sampler, Tempering, Non-homogeneous Markov Chain, Spectral Gap, State Decomposition, Cheeger's Inequality, Isoperimetric Inequality, Equi-Energy Sampler, Phylogenetic Tree

that a Small-World chain converges faster than a Metropolis-Hastings chain with local proposals when sampling a multi-modal density.

The new algorithm is simple and easy to implement. In its simplest form the algorithm has two Small-World chains: the first chain samples a fattened (or flattened) distribution (via tempering) with its long-range proposals generated by a typical heavy-tailed distribution; the second chain samples the target distribution with its long-range proposals randomly drawn from samples of the first chain. Intuitively, the first chain identifies and remembers (via its empirical distribution) modes of the distribution – it does so more effectively than an unheated chain because it accepts more long-range proposals that tend to move between different modes. This knowledge of whereabouts of different modes is used by the second chain through its long-range proposals.

The new algorithm bears similarity with the Equi-Energy sampler. Indeed, our original goal was to prove the ergodicity of the Equi-Energy sampler – the ergodicity proof in the Equi-Energy paper was incomplete and the amended proof Atchadé and Liu (2006) has difficulties. Identified these problems, Andrieu *et al.* (2008) proved ergodicity of the Equi-Energy sampler using the Poisson's Equation. By studying properties of perturbed kernels, Fort *et al.* (2010) proved ergodicity of a class of adaptive MCMC algorithms, including the Equi-Energy sampler. Our proof uses mixingale theory, built on the work of Atchadé and Liu (2006) and Atchadé and Rosenthal (2005), which allows us to study the asymptotic convergence rate. The convergence results identified that local proposal in the feeding chain of Equi-Energy sampler slows down the convergence in a multi-modal distribution. Thus we propose using a small-world sampler as the feeding chain.

The paper contributes both practically and theoretically in sampling high dimensional multi-modal distributions. First, we provide a simple and easy-to-implement algorithm that is effective for challenging problems. Second, we prove ergodicity of the algorithm and analyze its convergence rate. The result on convergence rate provides important insights into when, why and how the algorithm will improve convergence in practice.

We first formally set up the problem. Let $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ be the state space equipped with its σ-algebra. A Metropolis-Hastings algorithm (Metropolis *et al.*, 1953; Hastings, 1970) aims to sample from a probability measure π that admits a density with respect to Lebesgue measure that is only known up to a normalizing constant. We use π to denote a measure and $\pi(x)$ to denote a density and it should be clear from the context. The transition kernel of a Metropolis-Hastings chain is

$$P(x, dy) = a(x, y)\, k(x, dy) + r(x)\, \delta_x(dy), \tag{1}$$

where $k(x, y) : \mathbb{R}^d \times \mathbb{R}^d \to [0, \infty)$ is a *proposal distribution*, $a(x, y) = \min\left(1, \frac{\pi(y)}{\pi(x)} \frac{k(y,x)}{k(x,y)}\right)$ is the *acceptance probability* of a proposed move y, the δ_x is the point mass at x, and $r(x) = \int_{\mathbb{R}^d} (1 - a(x, y))k(x, dy)$ is the probability that y being rejected. Obviously $k(x, y)$ determines $P(x, dy)$.

Let $f : \mathbb{R}^d \to \mathbb{R}$ denotes a measurable function. For a signed measure μ and a positive function V, define the V-norm of μ by $||\mu||_V := \sup_{f \le V} |\mu(f)|$ where $\mu(f) = \int_{\mathbb{R}^d} f(x)\mu(dx)$. A transition kernel P acts on f such that $Pf(x) = \int_{\mathbb{R}^d} f(y)P(x, dy)$. Given two transition kernels P, Q, the

product PQ is also a transition kernel $(PQ)(x, A) = \int_{\mathbb{R}^d} P(x, dy)Q(y, A)$. This allows us to define a product of kernels of countable many through induction.

1.1. *Minorization and drift conditions.* We assume the following assumptions hold:

A1 A probability measure ψ exists on $\mathcal{B}$ such that the Markov chain is ψ-irreducible and aperiodic (c.f. Meyn and Tweedie, 1993).

A2 *Minorization Condition*: there exist a set $C \in \mathcal{B}$ and $\epsilon > 0$ such that, for the same ψ in A1, we have $\psi(C) > 0$ and, for all $A \in \mathcal{B}$, $x \in C$,

$$P(x, A) \geq \epsilon\, \psi(A). \tag{2}$$

A3 *Drift Condition*: there exist a measurable function $V : \mathbb{R}^d \to [1, \infty)$ such that $\pi(V) < \infty$ and constants $\lambda \in [0, 1)$ and $b \in (0, \infty)$ and same set C in A2 satisfying

$$PV(x) \leq \lambda V(x) + b\mathbf{1}_C(x). \tag{3}$$

The minorization condition ensures the *flow* from the set C to outside is lower-bounded, an idea intimately connected to conductance (see Section 4). The drift condition guarantees that the Markov chain evolves towards C. The drift condition is necessary to define V-geometrically ergodic. The above minorization and drift conditions can be checked for many practical problems. For example, if P is a Random Walk Metropolis kernel, both (2) and (3) are known to hold under some regularity conditions on the target densities (see Atchadé, 2010).

1.2. *Spectral gap.* A homogeous Markov chain that satisfies (A1-A3) is geometric ergodic (c.f. Meyn and Tweedie, 1993; Roberts and Rosenthal, 2004). Let $L^2(\pi)$ denote the space of measurable functions on $\mathbb{R}^d$ with $\int_{\mathbb{R}^d} f(x)^2\pi(dx) < \infty$, with inner product $\langle f, g\rangle = \int_{\mathbb{R}^d} f(x)g(x)\pi(dx)$, and norm $\|f\| = \langle f, f\rangle^{1/2}$. The operator P being reversible with respect to π is equivalent to P being self-adjoint. It is well known that the spectrum of P is a subset of $[-1, 1]$. (Self-adjoint implies its spectrum is real, and being a Markov transition kernel determines the range.) A chain is said to be $L^2(\pi)$-geometrically ergodic if there exist a constant $\rho \in (0, 1)$ and a positive $M < \infty$, and $V(x)$ defined in (A3) such that

$$\|P^n(x, \cdot) - \pi(\cdot)\|_V \leq M\rho^n V(x). \tag{4}$$

Define $L^2_0(\pi) = \{f \in L^2(\pi) : \langle f, \mathbf{1}\rangle = 0\}$. Denote by P_0 the restriction of P to $L^2_0(\pi)$. It has been shown (Roberts and Rosenthal, 1997; Roberts and Tweedie, 2001) that for reversible Markov chains, geometric ergodicity is equivalent to the condition

$$|||P_0||| \equiv \sup_{f\in L^2_0(\pi), \|f\|\leq 1} \|P_0 f\| < 1. \tag{5}$$

The *spectral gap* of the chain P is defined by

$$\mathbf{Gap}(P) = 1 - |||P_0|||. \tag{6}$$

The spectral gap determines the convergence speed of a MCMC algorithm. Very roughly, a chain is close to equilibrium after a few multiples of $1/\mathbf{Gap}(P)$ iterations (Madras and Randall, 2002).

1.3. *Piecewise log-concave distributions.* We assume the target density $\pi(x)$ is a mixture of log-concave densities. A function $f : \mathbb{R}^d \to (0, \infty)$ is *log-concave* if for any $s \in [0, 1]$,

$$f(s\,x + (1-s)y) \geq f(x)^s f(y)^{1-s}. \tag{7}$$

Let $|\cdot|$ be a metric on $\mathbb{R}^d$, f is *α-smooth* if $|\log f(x) - \log f(y)| < \alpha\,|x-y|$ for all $x, y \in \mathbb{R}^d$. Let $\{A_1, \ldots, A_m\}$ be a partition of $\mathbb{R}^d$. Let $\pi_{(i)} : A_i \to (0, \infty)$ be an α-smooth log-concave density with *barycenter* $\beta_i = \int_{A_i} \pi_{(i)}(dx)$ and the *first moment* $M_i = \int_{A_i} |x - \beta_i| \pi_{(i)}(dx)$. Let $\beta_{ij} = |\beta_i - \beta_j|,\ i \neq j$ and $\beta_\pi = \max(\beta_{ij})$. Let $w_i > 0$, the distribution of interest is

$$\pi(x) \propto \sum_{i=1}^{m} \pi_{(i)}(x) 1_{A_i}(x)\ w_i. \tag{8}$$

Define average proposal distance of $k(x, y)$ as $D = \int_{\mathbb{R}^d \times \mathbb{R}^d} |x - y| k(x, y) dx dy$. Then $k(x, y)$ is *local* if $D < \min\{M_i\}$ and *long-range* if $D > \beta_\pi$. We call a chain *local chain* if it only uses local proposals. By *size* (or *complexity*) of $\pi(x)$ we mean the quantities associated with $\pi(x)$ that will affect the spectral gap. Namely, the measure of the steepness of each mode α and the measure of the distances between modes β_{ij}. We treat dimensionality of $\mathbb{R}^d$ and number of modes m as fixed quantities.

1.4. *State decomposition theorem.* We describe the "pieces" of a Metropolis–Hastings chain P by defining, for each $i = 1, \ldots, m$, a new Markov chain on A_i that rejects any transitions of P out of A_i. The transition kernel P_{A_i} of the new chain is given by

$$P_{A_i}(x, B) = P(x, B) + 1_B(x) P(x, A_i^c) \quad \text{for } x \in A_i, B \subset A_i. \tag{9}$$

The movement of the original chain among the "pieces" can be modeled by a "component" Markov chain with state space $\{1, \ldots, m\}$ and transition probabilities:

$$P_c(i, j) = \frac{1}{2\,\pi(A_i)} \int_{A_i} P(x, A_j) \pi(dx), \quad \text{for } i \neq j, \tag{10}$$

and $P_c(i, i) = 1 - \sum_{j \neq i} P_c(i, j)$.

THEOREM 1.1 (State Decomposition Theorem). *In the preceding framework, as given by Equations (9) and (10), we have*

$$\mathbf{Gap}(P) \geq \frac{1}{2} \mathbf{Gap}(P_c)(\mathbf{min}_{i=1,\ldots,m} \mathbf{Gap}(P_{A_i})). \tag{11}$$

Guan and Krone (2007) proved this state decomposition theorem, generalized it from its original version (Madras and Randall, 2002). The theorem says that we can bound the spectral gap by taking into account of the mixing speed within each mode and that of among different modes. In Section 4, we will focus on $\mathbf{Gap}(P_c)$ because it causes slowly mixing.

1.5. *Tempering and Equi-Energy sampler.* Since Geyer (1991) and Marinari and Parisi (1992), *tempering* has become a popular technique. Here we use two chains to illustrate tempering, although it usually employs multiple chains (Geyer, 1991).

ALGORITHM 1.1. *Let $s \in (0,1)$. Let x_c and x_h be current states of the cold and hot chains that sample $\pi(x)$ and $\pi_t(x) \propto \pi(x)^{1/t}$ (for some $t > 1$) respectively. Repeat the following steps.*

1. *Simulate $u \sim$ Unif $[0,1)$.*
2. *If $u < s$, update x_c and x_h independently using the Metropolis-Hastings algorithm.*
3. *If $u \geq s$, compute $a = \frac{\pi_t(x_h)}{\pi_t(x_c)} \frac{\pi(x_c)}{\pi(x_h)}$ and swap x_c and x_h with probability min$(1,a)$.*

The hot chain samples a flattened distribution; the flattening makes it easier for a local chain to move around to discover new modes. This easiness in the hot chain transfers to the cold chain through successful "swap". However, tempering has at least two limitations. First, because the process is "memory-less", modes need to be repeatedly rediscovered. Second, the cold chain interferes with the hot chain because of the "swap", which may slow down mixing. The Equi-Energy sampler (Kou *et al.*, 2006) resolves both limitations using empirical distributions. The Equi-Energy sampler runs multiple chains of different temperatures and samples of each chain are recorded and classified into different energy rings according to their energy level (log density). In addition to local proposals, the Equi-Energy sampler has "equi-energy jumps", when a lower temperature chain proposes a new move by randomly drawing a sample from an energy ring in a higher temperature chain. Thus, the relative easiness of moving between different modes in hot chains propagates down to the cold chain. Note a high-temperature chain affects the proposal of a lower-temperature chain; while a lower-temperature chain does not affect a high-temperature chain.

The rest of the paper is organized as following. In the next section, we describe the algorithms and main results and compare the algorithm with the Equi-Energy sampler. In Section 3 we prove ergodic theorems and discuss the convergence rate. In Section 4 we prove theorems regarding spectral gaps of the algorithm. In Section 5 we discuss practical issues and applications. A short discussion will conclude the paper.

2. Algorithms and Main Results. The new algorithm combines a Small-World sampler with tempering and we call it "Small-world Tempering with Empirical Ensemble Propagation", or STEEP. It has a backronym "Small-world Tempering with Equi-Energy Program" to credit the Equi-Energy sampler. We begin with the Small-World sampler.

ALGORITHM 2.1 (Small World Sampler). *Let $s \in (0,1)$ and the current state is $X_n = x$. Let $l(x,y)$ and $h(x,y)$ be local and long-range proposals respectively.*

1. *With probability $(1-s)$ simulate y from density $l(x,y)$ and compute $a = \frac{\pi(y)}{\pi(x)} \frac{l(y,x)}{l(x,y)}$.*
2. *Otherwise, simulate y from density $h(x,y)$ and compute $a = \frac{\pi(y)}{\pi(x)} \frac{h(y,x)}{h(x,y)}$;*
3. *Set $X_{n+1} \leftarrow y$ with probability min$(1,a)$, otherwise set $X_{n+1} \leftarrow x$.*

4. *Set* $n \leftarrow n+1$, *goto step 1.*

When sampling a multi-modal density, both simulation studies (Guan *et al.*, 2006) and theoretical work (Guan and Krone, 2007) have shown that a Small-World sampler converges faster than a local chain. Intuitively, the local proposals allows the chain to better explore a mode, while the long-range proposals allows it to jump between modes more easily. However, if modes are steep and far apart, a Small-World sampler will fail because that successful transitions between modes (initiated by long-range proposals) depend on their volumes (Guan and Krone, 2007) – to increase volumes of modes, tempering does the trick.

2.1. *The STEEP Algorithm.* The simplest form of STEEP employs two chains that run simultaneously. Both are Small-World chains because their proposal distributions are mixtures of local and long-range. The *exploring chain* samples a fattened target $\pi_t(x) \propto \pi(x)^{1/t}$ for some $t > 1$, where the long-range proposals are drawn from a typical heavy-tailed distribution (e. g., Cauchy). Samples collected in the exploring chain induce an empirical distribution ξ. The *sampling chain* samples $\pi(x)$, using ξ to simulate its long-range proposals. Because the similarity between π_t and π, in particular, their modes are in the same places, the empirical distribution provides natural and "intelligent" long-range proposals for the sampling chain. In detail:

ALGORITHM 2.2 (STEEP with Two Chains). *Let* $\pi_t(x) \propto \pi(x)^{1/t}$ *for some* $t > 1$. $\{Y_n\}$ *and* $\{X_n\}$ *sample* π_t *and* π *respectively, and* $Y_n = y_n$ *and* $X_n = x_n$.

1. *Simulate* $Y_{n+1}|Y_n = y_n$ *using Algorithm 2.1 to obtain* y_{n+1}. *Update empirical measure* $\xi_{n+1} = n/(n+1)\xi_n + 1/(n+1)\delta(y_{n+1})$.
2. *Simulate* $X_{n+1}|X_n = x_n$ *using Algorithm 2.1. Modify step (2) so that* $y \sim \xi_{n+1}$ *and compute* $a = \frac{\pi(y)}{\pi(x_n)}\frac{\pi_t(x_n)}{\pi_t(y)}$. *Set* $X_{n+1} \leftarrow y$ *with probability min*$(1, a)$, *otherwise set* $X_{n+1} \leftarrow x_n$.
3. *Set* $n \leftarrow n+1$, *goto step 1.*

The algorithm 2.2 is similar to the algorithm in section 3.3 of Andrieu *et al.* (2007). The difference here is that we emphasize the long-range proposals.

REMARK 1 (Ergodicity). *The exploring chain is a homogenous Markov chain with stationary distribution* π_t. *Conditional on the exploring chain, the sampling chain is a non-homogeneous Markov chain, whose transition kernel evolves over time because it depends on* ξ_n. *Since* $\xi_n \to \pi_t$, *the sampling chain converges to a Small-World chain with long-range proposal* π_t *and stationary distribution* π. *As a result, if we run Algorithm 2.2 long enough the sampling chain should generate samples approximately from* π. *This intuition is formalized in the ergodic theorem in Section 3.*

Since the exploring chain samples a fattened distribution π_t, each mode is larger and hence more easily found by long-range proposals. This implies a better mixing compared to directly sampling π. We quantify the intuition in the following theorem.

THEOREM 2.1. *Denote E the exploring chain in Algorithm 2.2 with stationary distribution π_t, and denote E_c the component chain of E (see Section 1.4), then $\mathbf{Gap}(E_c) \geq c_1\ t^d$ for some constant c_1 that is independent of t and d.*

However, a large t increases dissimilarity between π_t and π. Large dissimilarity reduces acceptance ratio of long-range proposals in the sampling chain – causing slow mixing. We quantify this intuition in the following theorem.

THEOREM 2.2. *Denote S an (idealized) sampling chain in Algorithm 2.2, using $\pi_t(x)$ as its long-range proposal instead of ξ, and denote S_c the component chain of S (see Section 1.4), then $c_2 t^{-2d} \leq \mathbf{Gap}(S_c) \leq c_3\ t^{-d}$ for some constants c_2, c_3. In addition, $\mathbf{Gap}(S_c)$ is independent of the size of $\pi(x)$.*

REMARK 2 (Equilibrium Assumption). *Note that Theorem 2.2 considers a Small-World chain that uses π_t as its long-range proposal. The connection with Algorithm 2.2 is that, as the number of iterations increases, the long-range proposal ξ_n used by the Algorithm becomes an increasingly close approximation to π_t (indeed $\xi_n \to \pi_t$ weakly). Thus, intuitively, the result in the theorem represents the "asymptotic" behavior of the algorithm.*

Obviously, for Algorithm 2.2 to converge quickly, both the exploring chain E and the sampling chain S must converge quickly. Therefore, we want to increase $g = \min(\mathbf{Gap}(E_c), \mathbf{Gap}(S_c))$. Suppose we set $t = 1$ in Algorithm 2.2, then the (idealized) sampling chain converges instantaneously (because it proposes from the target distribution). We have $g = \mathbf{Gap}(E_c)$, which is spectral gap of a Small-World chain. When we increase t in Algorithm 2.2, we enlarge a small spectral gap, $\mathbf{Gap}(E_c)$, and pay a penalty by shrinking a large spectral gap, $\mathbf{Gap}(S_c)$, to achieve faster mixing. This trade-off works well because $\mathbf{Gap}(E_c)$ depends on target density and can be extremely small; while $\mathbf{Gap}(S_c)$ is independent of the target density. However, when d is large, even for a modest t, the penalty on $\mathbf{Gap}(S_c)$ can be large. This lead us to extend the algorithm to multiple chains, instead of just two.

ALGORITHM 2.3 (STEEP). *Let $\{t_i\}$ be a sequence of positive numbers with $1 = t_0 < t_1 < \cdots < t_H$. Let $\{X_n^{(i)}\}$ sample $\pi_i(x) \propto \pi(x)^{1/t_i}$. Let $\xi^{(i)}$ be the empirical measures by the i-th chain. We call the H-th chain* exploring *chain, the rest* sampling *chains.*

1. *Simulate $X_{n+1}^{(H)} | X_n^H = x_i^{(H)}$ with Algorithm 2.1 to obtain $x_{n+1}^{(H)}$. Update empirical measure $\xi_{n+1}^{(H)} = n/(n+1)\xi_n^{(H)} + 1/(n+1)\delta(X_{n+1}^{(H)})$.*
2. *For each $i = H-1, H-2, \ldots, 1, 0$:*
   - *Simulate $X_{n+1}^{(i)} | X_n^{(i)} = x_n^{(i)}$ using Algorithm 2.1 to obtain $x_{n+1}^{(i)}$. Modify step (2) so that $y \sim \xi_{n+1}^{(i+1)}$ and compute $a = \frac{\pi_i(y)}{\pi_i(x_n^{(i)})} \frac{\pi_{i+1}(x_n^{(i)})}{\pi_{i+1}(y)}$.*
   - *Set $X_{n+1}^{(i)} \leftarrow y$ with probability min$(1, a)$, otherwise $X_{n+1}^{(i)} \leftarrow x_n^{(i)}$.*

- *Update empirical measure* $\xi_{n+1}^{(i)} = n/(n+1)\xi_n^{(i)} + 1/(n+1)\delta(x_{n+1}^{(i)})$.

3. *Set* $n \leftarrow n+1$, *goto step 1.*

We may determine *optimal* temperatures $\{t_i\}$ in Algorithm 2.3. Assuming t_H fixed and all chains reach equilibrium, denote G_i a Markov chain that samples $\pi_i(x)$ with long-range proposal $\pi_{i+1}(x)$, and let g_i be the spectral gap of the component chain of G_i, for $i = 0, \cdots, H-1$. Theorems 2.2 implies that

$$g_0 = c_0(t_1/t_0)^{-d}, g_1 = c_1(t_2/t_1)^{-d}, \ldots, g_{H-1} = c_{H-1}(t_H/t_{H-1})^{-d}, \tag{12}$$

for some constants $c_0, \ldots, c_{H-1}$. We want to choose $\{t_i\}$ to optimize

$$g = \min(g_1, \ldots, g_{H-1}). \tag{13}$$

Assume $c_1 = \cdots = c_H$ in Equations (12), then for a fixed t_H, we have $g_0 = \cdots = g_{H-1}$ maximizes g defined in (13). This implies that $\{t_i\}$ is a geometric progression because of (12) and assumption on c_i's. Such a geometric progression on adjacent temperatures has been suggested in both parallel tempering (Predescu *et al.*, 2004) and the Equi-Energy sampler (Kou *et al.*, 2006). The former is based on the optimality of the acceptance ratio of swaps between adjacent chains and the later is based on empirical evidence. Our argument here provides an additional justification based on spectral gap.

The exploring chain in Algorithm 2.3 can benefit from a large t_H with a modest $\tau = t_{k+1}/t_k$ and a modest H thanks to the geometric progression. A large t_H makes the "exploring" easier, while the penalties spread out across sampling chains. We will discuss how to choose t_H and τ in practice in Section 5.

2.2. *Finding Global Optimum.* We may extend the STEEP to find the global optimum of a density $\pi(x)$ – tempering does the trick. Specifically, in Algorithm 2.3, instead of stop at $t = 1$, we continue the process at $t = \tau^{-k}$ for $k = 1, \cdots, C$. As k increases, each mode becomes steeper and the probability concentrates around the global maximum. The samples collected in the last chain will be very close to the global optimum. We have the following Lemma.

LEMMA 2.3. *Let* $t_k = \tau^k$ *for* $k = H, H-1, \cdots, 1, 0, -1, \ldots - C$, *then for a sufficiently large* C, *the samples collected by the last chain in Algorithm 2.3 will be arbitrarily close to the global maximum.*

PROOF. Let $\pi(x)$ be an α-smooth log-concave density that attains its unique maximum at 0. Assume $\pi_t(x) \propto \pi(x)^{1/t}$. It is sufficient to show that for any $\epsilon > 0$ there exists $\delta > 0$, such that for any $0 < t < \delta$ we have $\int_{B_\epsilon} \pi_t(x) > 1 - \epsilon$, where B_ϵ is a ball centered at 0 with radii ϵ. However, the claim hold trivially because $\pi_t(x)$ decays exponentially at rate α/δ which can be made arbitrarily large by letting $\delta \to 0$ (or equivalently increasing C). Extension to mixture of α-smooth log-concave densities is trivial if the mixture has a unique global optimum. If the mixture density

has multiple global optima, then each global optimum x_j has a ball $B_{j,\epsilon}$ centered around it and the samples collected will concentrate in $B_{j,\epsilon}$. And each global optimum can be found. □

2.3. *Connection with the Equi-Energy sampler.* Although our algorithm shares several similarities with (and benefits of) the Equi-Energy sampler, there are also several differences. First, our algorithm is conceptually simpler, effectively because it dispenses with the energy ring and energy cut-off of the Equi-Energy sampler. This makes it easier to understand, and perhaps also slightly easier to implement, which we believe could facilitate its more widespread use (despite its appeals, the Equi-Energy sampler appears not to have been used very extensively in practical applications). Second, from a more technical perspective, our algorithm makes use of a small-world (fast converging) chain as the exploring chain. This has two advantages. First, Theorem 2.1 gives us insights on why the tempering helps, while such a theorem does not hold for a local chain used in the Equi-Energy sampler. Second, it has been argued (Jarner and Roberts, 2007) that, if the target distribution π is heavy tailed, the proposal distributions should have heavy tails as well. When combining high temperature with energy cutoff, the target distribution is likely to behave like a heavy tailed distribution, which will be a challenge for a local chain, while not much so for a Small-World chain. In addition, STEEP provides a different perspective in that every chain in STEEP is a Small-World chain, and a hot chain provides informed long-range proposals for the adjacent cold chain. However, in practice, at least for the simple examples given in the end of the paper, we would not expect much difference in performance between STEEP and the Equi-Energy sampler.

3. Ergodicity. We devote this section to state and prove ergodicity, Theorem 3.5, for Algorithm 2.3 (multiple chains). We first extend a key result of (Atchadé and Rosenthal, 2005, Lemma 3.1) to two chains (Algorithm 2.2) to obtain Theorem 3.2. Then we extend a theorem in (Atchadé and Liu, 2006, Theorem 3.3) to V-geometric ergodic to obtain Theorem 3.4. Our proof follow closely to theirs. There are two technical lemmas: Lemma 3.1 contributes to prove Theorem 3.2; and Lemma 3.3 helps Theorem 3.2 to prove Theorem 3.4. We then state and prove Theorem 3.5. We conclude this section by discussing convergence rate of Algorithms 2.2 and 2.3.

In Algorithm 2.2 $\{Y_n\}$ is the exploring chain and $\{X_n\}$ is the sampling chain. Let $X_0 = x_0$, $Y_0 = y_0$. Let $\{E_n\}$ denote a sequence of operators that generates $\{Y_n\}$, where E_i may be identical. Denote $E_{1:n} = E_1 \cdots E_n$. Recall ξ_n is an empirical measure generated by process $\{Y_n\}$. Let $\{P_{\xi_n}\}$ denote the sequence of operators that generates $\{X_n\}$, where

$$P_{\xi_n}(x, A) = (1 - s)T(x, A) + sK_{\xi_n}(x, A), \tag{14}$$

where T is a local chain with stationary distribution π and

$$K_{\xi_n} f(x) = \frac{1}{n}\sum_{j=1}^{n} \alpha(x, y_j) f(y_j) + \frac{1}{n} f(x) \sum_{j=1}^{n} (1 - \alpha(x, y_j)), \tag{15}$$

where y_i's are samples that induce empirical measure ξ_n, and $\alpha(x,y) = \min\left(1, \frac{\pi(y)}{\pi(x)}\frac{\pi_t(x)}{\pi_t(y)}\right)$. For a sequence of operators $\{P_n\}$, denote $P_{i:j} = P_i P_{i+1} \cdots P_j$ for $i < j$, $P_{i:i} = P_i$, and $P_{i:j} = I$ if $i > j$. Let $\mathcal{F}_n^X = \sigma(X_1, \ldots, X_n)$ and $\mathcal{F}_n^Y = \sigma(Y_1, \ldots, Y_n)$ be the filtrations of process $\{X_n\}$ and $\{Y_n\}$ respectively, and let $\mathcal{F}_n = \sigma(X_1, \ldots, X_n, Y_1, \ldots, Y_n)$.

The following Lemma is needed to prove Theorem 3.2.

LEMMA 3.1. *Define $\alpha_n := ||\xi_n - \xi_{n-1}||_V$, and assume there exist constants $M < \infty$ and $\rho \in (0,1)$, such that for each $x \in \mathcal{X}$ $||P_{\xi_n}^j(x,\cdot) - \pi_{\xi_n}(\cdot)||_V \leq M\rho^j V(x)$ P_{y_0}-a.s., then*

1. *α_n is $\mathcal{F}_n^Y$ measurable and $\mathbb{E}_{y_0}(\alpha_n) \leq O(1/n)$.*
2. *$||P_{\xi_n}(x,\cdot) - P_{\xi_{n-1}}(x,\cdot)||_V < 2s\ V(x) \cdot \alpha_n$.*
3. *$||\pi_{\xi_n} - \pi_{\xi_{n-1}}||_V \leq M\alpha_n$ for some constant M.*

PROOF. $\alpha_n = \sup_{|f|<V} |\xi_n(f) - \xi_{n-1}(f)| \leq \frac{1}{n}|V(Y_n) + \frac{1}{n-1}\sum_{i=1}^{n-1} V(Y_i)|$ and note that $\mathbb{E}_{y_0}(V(Y_n)) < \infty$ for all n and that $\mathbb{E}_{y_0}(V(Y_n)) \to \mathbb{E}(V) < \infty$ as $n \to \infty$, and claim (1) follows. Let $|f| < V$, then

$$
\begin{aligned}
&|P_{\xi_n}(x,f) - P_{\xi_{n-1}}(x,f)| \\
&\quad = s\left|\int \alpha(x,y)f(y)[\xi_n - \xi_{n-1}](dy) + f(x)\int [1-\alpha(x,y)][\xi_n - \xi_{n-1}](dy)\right| \\
&\quad \leq s\left|\int [f(y) + f(x)][\xi_n - \xi_{n-1}](dy)\right| \qquad (16)\\
&\quad \leq 2\ sV(x)\left|\int V(y)[\xi_n - \xi_{n-1}](dy)\right| \\
&\quad \leq 2\ sV(x)||\xi_n - \xi_{n-1}||_V,
\end{aligned}
$$

and the conclusion (2) follows. From Lemma B.1. of Andrieu *et al.* (2007), we have

$$
|(P_{\xi_n}^k - P_{\xi_{n-1}}^k)(x,f)| \leq M \sup_{x\in\mathbb{R}^d} \frac{||(P_{\xi_n}(x,\cdot) - P_{\xi_{n-1}}(x,\cdot)||_V}{V(x)},
$$

for all k. Let $k \to \infty$ and combine with (2) to get desired result (3). □

THEOREM 3.2. *In the proceeding framework, and assume*

1. *For each $x \in \mathcal{X}$, $||E_{1:n}(x,\cdot) - \pi_t(\cdot)||_V \to 0$, as $n \to \infty$.*
2. *There exist constants $M < \infty$ and $\rho \in (0,1)$, such that for each $x \in \mathcal{X}$ $||P_{\xi_n}^j(x,\cdot) - \pi_{\xi_n}(\cdot)||_V \leq M\rho^j V(x)$ P_{y_0}-a.s.*

In addition, for finite constant c_1, c_2, define

$$
B(c_1, c_2, n) = min_{1\leq k\leq n}[c_1\frac{k}{n-k} + c_2\rho^k]. \qquad (17)
$$

Define $g_{k,\xi_k} = f - \pi_{\xi_k}(f)$. *Then for any* $|f| \le V$, *we have*

$$|\mathbb{E}_{(x_0,y_0)}(g_{n+j,\,\xi_{n+j}}(X_{n+j})\big|\mathcal{F}_n)| \le B(k_1,k_2,j)V(X_n) \tag{18}$$

P_{x_0,y_0} *a.s. and as an immediate consequence,*

$$|\mathbb{E}_{(x_0,y_0)}(f(X_n) - \pi_{\xi_n}(f))\,| \le B(k_1,k_2,n)V(x_0). \tag{19}$$

In addition,

$$\frac{1}{n}\sum_{i=1}^{n}(f(X_i) - \pi_{\xi_i}(f)) \to 0 \quad \textit{as } n \to \infty, P_{x_0,y_0} - a.s. \tag{20}$$

PROOF OF THEOREM 3.2. For notational convenience, we denote P_{ξ_n} by P_n in the following equation.

$$\begin{aligned}
&\left|\mathbb{E}_{x_0}(g_{n,\,\xi_n}(X_{n+j})|\mathcal{F}_n)\right| = \left|P_{n:(n+j-1)}f(X_n) - \pi_{\xi_n}(f)\right| \\
&\le \Big|\sum_{k=1}^{j-1}(P_n^k - \pi_{\xi_n})(P_{n+k} - P_n)P_{(n+k+1):(n+j-1)}f(X_n)\Big| + \left|P_n^j f(X_n) - \pi_{\xi_n}(f)\right| \\
&\le \Big|\sum_{k=1}^{j-1} M_1(P_n^k - \pi_{\xi_n})(P_{n+k} - P_n)V(X_n)\Big| + M_2\rho^j V(X_n) \\
&\le \sum_{k=1}^{j-1} M_3 k\alpha_n(P_n^k - \pi_{\xi_n})V(X_n) + M_2\rho^j V(X_n) \\
&\le M_4\alpha_n V(X_n)\sum_{k=1}^{j-1}\rho^k k + M_2\rho^j V(X_n) \\
&\le [M_4\frac{\alpha_n}{(1-\rho)^2} + M_2\rho^j]V(X_n),
\end{aligned}$$

where α_n is defined in Lemma 3.1 (1). The transition between the first and the second line comes from the well known identity $P_{1:n} = \sum_{k=1}^{n-1} P_1^k(P_{k+1} - P_1)P_{(k+2):n} + P_1^n$; in the transition between the second and the third line, we recursively apply drift condition (A3); and in the transition between the third and the fourth line we applied Lemma 3.1 (2) with telescoping sum. Now taking into account of Lemma 3.1 (3) and again with telescoping sum, we have

$$\begin{aligned}
&\left|\mathbb{E}_{x_0}(g_{n+j,\,\xi_{n+j}}(X_{n+j})|\mathcal{F}_n)\right| < [M_4\frac{\alpha_n}{(1-\rho)^2} + M_2\rho^j]V(X_n) + jM\alpha_n \\
&< [M_5\alpha_n\, j + M_2\rho^j]V(X_n),
\end{aligned} \tag{21}$$

where $M_5 = 2\max(M_4/(1-\rho)^2, M)$. Using the filtration trick in the end of Lemma 3.1 of Atchadé and Rosenthal (2005), we get for $k = 1, \ldots, j$

$$\begin{aligned}
&\left|\mathbb{E}_{x_0}(g_{n+j,\,\xi_{n+j}}(X_{n+j})|\mathcal{F}_n)\right| = \left|\mathbb{E}_{x_0}\left[\mathbb{E}_{x_0}\left(g_{n+j,\,\xi_{n+j}}(X_{n+j})\middle|\mathcal{F}_{n+j-k}\right)\middle|\mathcal{F}_n\right]\right| \\
&\leq \mathbb{E}_{x_0}\left[\left|\mathbb{E}_{x_0}\left(g_{n+j,\,\xi_{n+j}}(X_{n+j})\middle|\mathcal{F}_{n+j-k}\right)\right|\middle|\mathcal{F}_n\right] \\
&\leq \min_{1\leq k\leq j}[M_5\alpha_{n+j-k}\ k + M_2\rho^k]\mathbb{E}_{x_0}(V(X_n+j-k)|\mathcal{F}_n) \\
&\leq \min_{1\leq k\leq j}[M_5\alpha_{n+j-k}\ k + M_2\rho^k]V(X_n).
\end{aligned}$$

Taking expectation with respect to process $\{Y_n\}$, we get

$$\begin{aligned}
&|\mathbb{E}_{x_0,y_0}(g_{n+j,\,\xi_{n+j}}(X_{n+j})|\mathcal{F}_n)| \leq \mathbb{E}_{y_0}\{\min_{1\leq k\leq j}[M_5\alpha_{n+j-k}\ k + M_2\rho^j]\}V(X_n) \\
&\leq \min_{1\leq k\leq j}[M_6\frac{k}{n+j-k} + M_2\rho^j]V(X_n),
\end{aligned} \tag{22}$$

which is claim (18). Taking $n = 0$ in (22), we obtain (19).

Note $B(c_1, c_2, n) \to 0$ as $n \to \infty$ (see Section 3.1), this and (22) show that

$$\mathbb{E}_{x_0,y_0}(f(X_n) - \pi_{\xi_n}(f)) \to 0, \text{ as } n \to \infty. \tag{23}$$

Following argument similar to the proof of Theorem 3.2 of Atchadé and Rosenthal (2005), write $Z_n = g_{n,\xi_n}(X_n) - \mathbb{E}_{x_0,y_0}(g_{n,\xi_n}(X_n))$. Given (22), $(Z_n, \mathcal{F}_n)$ is an L^2-mixingale with mixingale sequence c_n being a constant and $\psi_n = B(c_1, c_2, n)$. Then Corollary 2.1 of Davidson and de Jong (1997) implies that

$$\frac{1}{n}\sum_{k=1}^{n}(g_{k,\xi_k}(X_k) - \mathbb{E}_{x_0,y_0}g_{k,\xi_k}(X_k)) \to 0, \quad P_{x_0,y_0} \text{ a.s. as } n \to \infty. \tag{24}$$

Combine (23) and (24) we obtain (20). □

The following lemma is needed to prove the Theorem 3.4. It is a technical lemma that is an extension of of Lemma 3.1 in Atchadé and Liu (2006).

LEMMA 3.3. *Let $\{f_n\}$ be a sequence of measurable functions and let $\{\mu_n\}$ be a sequence of probability measures such that $|f_n| < V$ and $f_n \to f$ pointwise and $\mu_n \to \mu$ setwise. In addition, $\int V(x)\mu_n(dx) < \infty$ and $\int V(x)\mu(dx) < \infty$. Then $\int f_n(x)\mu_n(dx) \to \int f(x)\mu(dx)$.*

PROOF OF LEMMA 3.3. In light of Proposition 18 in Chapter 11 of Royden (1988), it is sufficient to prove that $\int V(x)\mu_n(dx) \to \int V(x)\mu(dx)$. However, setwise convergence of $\mu_n \to \mu$ implies that for any simple function ψ, we have $\int \psi\mu_n(dx) \to \int \psi\mu(dx)$. Since simple functions are dense (in L^p), we have two sequences of simple functions $2V > \psi_m \geq V$ and $\phi_m \leq V$ such that $\psi_m \searrow V$ and $\phi_m \nearrow V$, and $\int \psi_m\mu_n(dx) \geq \int V\mu_n(dx) \geq \int \phi_m\mu_n(dx)$. Let $n \to \infty$, we have $\int \psi_m\mu(dx) \geq \lim_n \int V\mu_n(dx) \geq \int \phi_m\mu(dx)$ for any m. Now let $m \to \infty$ and apply Lebesgue's dominant convergence theorem to finish the proof. □

THEOREM 3.4. *With the setting outlined as in Theorem 3.2, assume:*

1. *For any $x \in \mathbb{R}^d$ and $A \in \mathcal{B}$, $P_{\xi_n}(x, A) \to P_\xi(x, A)$ as $n \to \infty$, P_{y_0}-a.s.*
2. *There exists a finite constant M and a $\rho \in (0, 1)$ such that $||P^k_\xi(x, \cdot) - \pi(\cdot)||_V < MV(x)\rho^k$ and $||P^k_{\xi_n}(x, \cdot) - \pi_{\xi_n}(\cdot)||_V < MV(x)\rho^k$ P_{y_0}-a.s.*

Then for any measurable function $f : \mathcal{X} \to \mathcal{R}$ such that $|f| < V$ we have

$$\mathbb{E}_{x_0,y_0}[f(X_n)] \to \pi(f) \quad \text{and} \quad \frac{1}{n}\sum_{k=1}^{n} f(X_k) \to \pi(f) \ \ P_{x_0,y_0} - a.s. \tag{25}$$

PROOF OF THEOREM 3.4. For any $|f| < V$, by Theorem 3.2 we have $\mathbb{E}_{x_0,y_0}[f(X_n) - \pi_{\xi_n}(f)] \to 0$ and $\frac{1}{n}\sum_{k=1}^{n}[f(X_k) - \pi_{\xi_k}(f)] \to 0$ P_{x_0,y_0}-a.s. as $n \to \infty$. To finish we need to prove that $\pi_{\xi_n}(f) \to \pi(f)$ P_{y_0}-a.s. as $n \to \infty$. By assumption we have $P_{\xi_n} f(x) \to P_\xi f(x)$ P_{y_0}-a.s. for all $x \in \mathbb{R}^d$. By Lemma 3.3, $P^2_{\xi_n} f(x) = P_{\xi_n}(P_{\xi_n} f(x)) \to P^2_\xi f(x)$. By recursion, for any $k \geq 1$,

$$P^k_{\xi_n} f(x) \to P^k_\xi f(x) \quad P_{y_0}\text{-a.s. as } n \to \infty. \tag{26}$$

We have

$$\begin{aligned} |\pi_{\xi_n}(f) - \pi(f)| &\leq |\pi_{\xi_n}(f) - P^k_{\xi_n} f(x)| + |P^k_\xi f(x) - \pi(f)| + |P^k_{\xi_n} f(x) - P^k_\xi f(x)| \\ &\leq 2MV(x)\rho^k + |P^k_{\xi_n} f(x) - P^k_\xi f(x)|. \end{aligned} \tag{27}$$

Combine above with (26) we have $|\pi_{\xi_n}(f) - \pi(f)| \to 0$ P_{y_0}-a.s. □

Recall in Algorithm 2.3, the H-th chain is a homogeneous chain. Conditioning on the realization of $\{X^{(i+1)}_n\}$ (for $i < H$), $\{X^{(i)}_n\}$ is a nonhomogeneous Markov chain with transition kernel $P^{(i)}_n$. The $P^{(i)}_n$ operates on f such that: $P^{(i)}_n f(x) = (1 - s)T^{(i)} f(x) + sK^{(i)}_{\xi^{(i+1)}_n} f(x)$, where $T^{(i)}$ is a homogeneous local chain with stationary distribution π_i, and

$$K^{(i)}_{\xi^{(i+1)}_n} f(x) = \frac{1}{n}\sum_{j=1}^{n} \alpha(x, y_j) f(y_j) + \frac{1}{n} f(x) \sum_{j=1}^{n} (1 - \alpha(x, y_j)), \tag{28}$$

where y_i's are samples that induce empirical measure $\xi^{(i+1)}_n$, and $\alpha(x, y) = \min\left(1, \frac{\pi_i(y)}{\pi_i(x)} \frac{\pi_{i+1}(x)}{\pi_{i+1}(y)}\right)$. Let $P^{(i)}(x, A) = (1-s)T^{(i)} f(x) + sK^{(i)}_{\xi^{(i+1)}} f(x)$ be the limiting transition kernel as $n \to \infty$ where

$$K^{(i)}_{\xi^{(i+1)}} f(x) = \int_{\mathbb{R}^d} \alpha(x, y) f(y) \pi_{i+1}(dy) + f(x) \int_{\mathbb{R}^d} (1 - \alpha(x, y))\, \pi_{i+1}(dy).$$

We have the following ergodic theorem.

THEOREM 3.5. *In proceeding framework, assume $T^{(i)}, i \in \{0, \ldots, H\}$, satisfies assumptions (A1-A3), then for $|f| < V$, as $n \to \infty$,*

$$\mathbb{E}[f(X_n^{(i)})] \to \pi_i(f) \quad \textit{and} \quad \frac{1}{n}\sum_{k=1}^{n} f(X_k^{(i)}) \to \pi_i(f) \quad a.s. \tag{29}$$

PROOF OF THEOREM 3.5. Let $x_0^{(i)}$ be the starting point of process $\{X_n^{(i)}\}$. It is easy to check the detailed blance equation holds for the process $\{X_n^{(H)}\}$, so the claim holds for chain H.

For each $i < H$, we want to show that the i-th chain in STEEP algorithm satisfies two assumptions in Theorem 3.4.

Since process $\{X_n^{(i+1)}\}$ has stationary distribution π_{i+1}, we have for all $x \in \mathbb{R}^d$, $P_n^{(i+1)}(x, \cdot) \to \pi_{i+1}$. Since $\alpha(x, y)$ as in (28) is bounded, by Lebesgue's dominate convergence theorem, $K^{(i)}_{\xi_n^{(i+1)}} f(x) \to K^{(i)}_{\xi^{(i+1)}} f(x)$, $P_{x_0^{(i+1)}}$-a.s., which implies $P_n^{(i)} f(x) \to (1-s)T^{(i)} f(x) + sK^{(i)}_{\xi^{(i+1)}} f(x)$ $P_{x_0^{(i+1)}}$-a.s., and hence $P_n^{(i)}(x, A) \to P^{(i)}(x, A)$ $P_{x_0^{(i+1)}}$-a.s. for each $x \in \mathbb{R}^d$ and $A \in \mathcal{B}$. So the Assumption (1) hold true.

The drift and minorization conditions on $T^{(i)}$ transfers to $P_n^{(i)}$, so that each $P_n^{(i)}$ admits an invariant distribution $\pi_{i,n}$ $P_{x_0^{(i+1)}}$-a.s. and is geometric ergodic. The limiting transition kernel $P^{(i)}$ also inherit the drift and minorization conditions on $T^{(i)}$ and admits invariant distribution π_i $P_{x_0^{(i+1)}}$-a.s. So the Assumption (2) hold true.

By induction, Theorem 3.5 now follows Theorem 3.4. □

3.1. *Convergence Rate.* The convergence of Algorithm 2.2 was broken down into two parts. The first part is $(P_{\xi_1} \cdots P_{\xi_n})(x, \cdot) \to \pi_{\xi_n}$ (Theorem 3.2), which is the convergence of the sampling chain; and the second part is $\pi_{\xi_n} \to \pi_t$ (Theorem 3.4), which is the convergence of the exploring chain. We shall discuss them separately.

First note the exploring chain in Theorem 3.2 is more general in that it can be either a homogenous or a non-homogenous chain. From (27) and Lemma 3.1

$$\begin{aligned} |\pi_{\xi_n}(f) - \pi(f)| \leq & 2MV(x)\rho^k + |P_{\xi_n}^k f(x) - P_{\xi}^k f(x)| \\ \leq & 2MV(x)\rho^k + 2sV(x)\|\xi_n - \xi\|_V \\ \leq & (c_1\rho^k + c_2\|\xi_n - \xi\|_V)V(x), \end{aligned} \tag{30}$$

where for any $|f| < V$, we have $\xi(f) = \pi_t(f)$ P_{y_0}-a.s. Clearly, the convergence rate of $\pi_{\xi_n} \to \pi$ is dominated by the rate $c_2\|\xi_n - \xi\|_V$, which depends on the convergence rate of the exploring chain.

In light of the discussion in Atchadé and Rosenthal (2005), Theorem 3.2 implies that the sampling chain in Algorithm 2.2 converges (to π_{ξ_n}) at rate

$$B(n, \rho) = \min_{1 \leq k \leq n}(c_1 k/(n-k) + c_2\rho^k). \tag{31}$$

Take derivative with respect to k and set to 0 to get

$$c_1 \frac{n}{(n-k)^2} = c_2 \log \frac{1}{\rho}\, \rho^k. \tag{32}$$

If we assume $k = O(n)$, then (32) reduce to $\frac{1}{n} \approx c\rho^k$, solve to get $k = O(\log(n))$, and we reach contradiction. So we may assume $k = o(n)$, and (32) simplifies to $c_1 \frac{1}{n} \approx c_2 \log \frac{1}{\rho}\, \rho^k$. Take log on both sides and solve to get $k \approx -\log n / \log \rho$. Substitute back to (31) and use $\log(1-x) \approx -x$ when x small and note $\rho^k \approx 1/n$ we get

$$B(n, \rho) \approx O(\frac{n^{-1} \log n}{1-\rho}). \tag{33}$$

Loosely, $(1-\rho) \approx Gap(P_\xi)$, where P_ξ is the limiting transition kernel of the sampling chain. From (33) the sampling chain of Algorithm 2.2 converges at a polynomial rate that is also proportional to $1/\mathbf{Gap}(P_\xi)$. Note the rate is not a function of the *size* of the $\pi(x)$.

Extending to Algorithm 2.3, the above analysis suggests that the sampling chains converge at rate $O(\tau^{\lambda d}\ n^{-1} \log n)$ for some $\lambda \in [1, 2]$ (Theorem 2.2), where $\tau > 1$, is the ratio between adjacent temperatures. A large τ slows down the convergence of the sampling chains. However, the spectral gap of the exploring chain is enlarged by a factor of τ^{Hd} (Theorem 2.1). Hard problems will benefit from such a trade-off because their exploring chains usually have very small spectral gaps.

4. Spectral Gaps. Our main aim in this section is to prove Theorems 2.1 and 2.2. We rely on the state decomposition theorem (Theorem 1.1) to analyze the Algorithm 2.2: we partition a multi-modal distribution into pieces log-concave pieces and analyze each restricted local chain P_{A_i} and the component chain P_c separately. The $\mathbf{Gap}(P_{A_i})$ is well studied (Lovász and Vempala, 2003; Mathé and Novak, 2007; Rudolf, 2009; Guan and Krone, 2007) using the isoperimetric inequality (Lovász and Simonovits, 1993; Kannan *et al.*, 1995), and $\mathbf{Gap}(P_c)$ can be computed directly using conductance and the Cheeger's inequality.

4.1. *Conductance and spectral gap.* Recall $P(x, dy)$ defined in (1), for $A \in \mathcal{B}$ with $\pi(A) > 0$, define

$$\mathbf{h}_P(A) = \frac{1}{\pi(A)} \int_A P(x, A^c) \pi(dx). \tag{34}$$

The quantity $\mathbf{h}_P(A)$ can be thought of as the (probability) flow out of the set A in one step when the Markov chain is at stationarity. The *conductance* of the chain is defined by

$$\mathbf{h}_P = \inf_{0 < \pi(A) \le 1/2} \mathbf{h}_P(A). \tag{35}$$

Intuitively, a small $\mathbf{h}_P$ implies mixing slowly because the chain may be trapped in a set whose measure is $\le 1/2$. On the other hand, a large $\mathbf{h}_P$ implies mixing rapidly as nowhere is sticky. We have the following theorem (Lawler and Sokal, 1988).

THEOREM 4.1 (Cheeger's Inequality). *Let P be a reversible Markov transition kernel with invariant measure π. Then*

$$\frac{\mathbf{h}_P^2}{2} \leq \mathbf{Gap}(P) \leq 2\,\mathbf{h}_P. \tag{36}$$

Let $k(x, dy) = (1-s)k_1(x, dy) + s\,k_2(x, dy)$, with $0 \leq s \leq 1$. Suppose operators P, P_1, and P_2 are induced by k, k_1, k_2 respectively. Clearly,

$$P = (1-s)P_1 + s\,P_2. \tag{37}$$

It is straightforward to show the following Lemma (Guan and Krone, 2007), which allows one to bound the spectral gap for a mixture of kernels.

LEMMA 4.2. *Suppose P is defined by (37). Then*

$$\mathbf{Gap}(P) \geq \frac{1}{2}\max\left((1-s)^2\,\mathbf{h}_{P_1}^2, s^2\mathbf{h}_{P_2}^2\right); \tag{38}$$

The following theorem was proved in Guan and Krone (2007).

THEOREM 4.3. *Suppose $\pi(x)$ is an α-smooth log-concave probability density of d-dimension on a convex set K. Suppose further that π has barycenter 0 and set $M_\pi = \int_K |x|\,\pi(dx)$. Then the conductance, $\mathbf{h}_P$, of the Metropolis-Hastings chain with transition kernel $P(x, dy)$ induced by the uniform δ-ball proposal satisfies*

$$\mathbf{h}_P \geq \frac{\delta\, e^{-\alpha\,\delta}}{1024\,\sqrt{d}\,M_\pi}. \tag{39}$$

Combining Equations (38) and (39) we obtain a lower bound of the spectral gap of a local chain sampling a log-concave distribution. Set $\delta = 1/\alpha$ to see that it is fast-mixing. This and the state decomposition theorem leads to a conclusion that a Small-World chain is fast mixing (Guan and Krone, 2007). Moreover, because the dimensionality d is in a polynomial form in (39), so there is no "curse of dimensionality" with a local chain sampling a log-concave distribution. On the contrary, there is a "curse of dimensionality" in the component chain simply because "volumes" of modes matter (Guan and Krone, 2007). This justifies our focus on the component chain in Theorems 2.1 and 2.2.

4.2. *Proof of The Thorems 2.1 and 2.2.* We need to establish the connection between a log-concave distribution and its powered-up alternatives.

LEMMA 4.4. *Let $f(x)$ be a log-concave distribution on $\mathbb{R}^d$ of dimension d, and $f(x)$ attain its unique maximum at 0. Let $f_t(x) \propto f(x)^{1/t}$, where $t > 1$. Then (a) $\frac{f_t(0)}{f(0)} \geq t^{-d}$, (b) $\frac{f_t(x)}{f(x)} \geq \frac{f_t(0)}{f(0)}$ for any $x \in \mathbb{R}^d$, and (c) if $f(x)$ is α-smooth, then the equality in (a) and (b) holds up to a constant.*

PROOF. By definition of log-concave, for $0 \le s \le 1$,

$$f(s\,x + (1-s)y) \ge f(x)^s f(y)^{1-s}.$$

Let $s = 1/t$ and $y = 0$. Choose h, such that $1/t + 1/h = 1$, then

$$f(x/t) \ge f(x)^{1/t} f(0)^{1/h}. \tag{40}$$

Since $\int_{\mathbb{R}^d} f(x)dx = 1$, we can get $\int_{\mathbb{R}^d} f(x/t)dx = t^d \int_{\mathbb{R}^d} f(x/t)d(x/t) = t^d$. So

$$\int_{\mathbb{R}^d} f(x)^{1/t} f(0)^{1/h} dx = f(0)^{1/h} \int_{\mathbb{R}^d} f(x)^{1/t} dx \le t^d,$$

rearrange term to get: $\frac{f(0)^{1/t}}{\int_{\mathbb{R}^d} f(t)^{1/t} dx} \ge f(0)\; t^{-d}$. Identify the left hand side is $f_t(0)$ to prove (a).

Since $f(x)$ is log-concave, for any unit vector $u \in \mathbb{R}^d$, $f(uz)$ is monotone decreasing in z, hence $f(zu)^{1/t-1}$ is monotone increasing in z (as we assume $t > 1$). This proves (b).

Because of logconcavity, for any unit vector $u \in \mathbb{R}^d$, there exists scalars g, ν, such that $f(uz) < e^{-\nu z}$ for $z > g$. By α-smooth, we have $f(uz) > e^{-\alpha z}$. For exponential functions, equality in (40) holds and this proves (c). □

REMARK 3. *Note the bound in (a) is essentially tight for a general log-concave distribution, as can be seen clearly from one-dimensional exponential distributions. For distributions such as (multivariate) normal, the bound in (a) is not tight. However, we are confined by the technical condition of α-smoothness, and we do not pursue a better bound. Although the α-smooth is a technical condition for the convenience to bound conductance (Guan and Krone, 2007), it is worthwhile to note that it is crucial for the tempering scheme as well. Taking an extreme example, the tempering will not help for two uniform distributions on two unit discs that are far apart. While tempering is helpful for two normal distributions that are far apart, as we will see in Section 5.*

LEMMA 4.5. *For a piece-wise α-smooth log-concave distributions defined in (8), let $\xi_i(x) = w_i^{1/t}\pi_i(x)^{1/t}/I_i$, where $I_i = w_i^{1/t}\int_{A_i}\pi_i(x)^{1/t}dx$ for $i = 1, \ldots, m$. Let $I = \sum I_i/m$ and $\xi_i'(x) = w_i^{1/t}\pi_i(x)^{1/t}/I$ for each i. Then there exists constants c_1, c_2 such that $c_1 < \xi_i'(A_i)/\xi_i(A_i) < c_2$.*

PROOF. For any pair $w_i \ge w_j$, for $t \ge 1$, we have $w_j/w_i \le w_i^{1/t}/w_j^{1/t} \le w_i/w_j$. So that without loss of generality, we may assume $w_i = 1$ for $i = 1, \ldots, m$. By (c) of Lemma 4.4, we have $\xi_i(A_i) = c_i\pi_i(A_i)t^{-d}/I_i$. Since $\xi_i(A_i) = \pi_i(A_i) = 1$, we have $I_i = c_i t^{-d}$, which implies that $\xi_i'(A_i)/\xi_i(A_i) = m\; c_i/\sum c_i$ and the conclusion follows. □

REMARK 4. *Lemma 4.5 says two different normalizations, namely, normalization within each pieces and normalization combining all pieces, are equivalent up to a constant for a piece-wise α-smooth log-concave distribution. So we can use the normalization within each pieces to ease the presentation.*

With Lemma 4.4 at hand, the proof of Theorem 2.2 is easy.

PROOF OF THEOREM 2.2. Consider π_{A_1}, π_{A_2} as the π restricted on A_1 and A_2 respectively, which we denote by π_1 and π_2. Let ξ_1, ξ_2 be their normalized powered-up alternatives respectively. The conductance bound of P_c requires estimate of the integral appeared in (10)

$$\int_{A_1\times A_2} \min\left(1, \frac{\pi_2(y)}{\pi_1(x)}\frac{\xi_1(x)}{\xi_2(y)}\right)\pi_1(x)\xi_2(y)dxdy \tag{41}$$

$$= \int_{A_1\times A_2} \min\left(\frac{\xi_1(x)}{\pi_1(x)}, \frac{\xi_2(y)}{\pi_2(y)}\right)\pi_1(x)\pi_2(y)dxdy \tag{42}$$

$$= c_1\tfrac{1}{t^d}\pi_1(A_1)\pi_2(A_2), \tag{43}$$

where the last equality obtained from Lemma 4.4. Therefore, from Equation (35) we get for some constant c

$$h_{12} = \frac{1}{2\pi_1(A_1)}\int_{A_1\times A_2} \min\left(1, \frac{\pi_2(y)}{\pi_1(x)}\frac{\xi_1(x)}{\xi_2(y)}\right)\pi_1(x)\xi_2(y)dx = \frac{c}{t^d}. \tag{44}$$

Hence, the conductance of the component chain P_c is proportional to t^{-d}. Following the Cheeger's Inequality (36) to get the bound on spectral gap. It is clear that the bound is *not* a function of the "size" of π. □

PROOF OF THEOREM 2.1. Use same notations defined in the proof of Theorem 2.2. Let $h(x, y)$ be the long range proposal. We have

$$\begin{aligned} h_{12} &= \frac{1}{2\pi_1(A_1)}\int_{A_1\times A_2} \min\left(1, \frac{\pi_2(y)}{\pi_1(x)}\frac{h(y,x)}{h(x,y)}\right)\pi_1(x)h(x,y)dxdy \\ &= \frac{1}{2\pi_1(A_1)}\int_{A_1\times A_2} \min\left(\frac{h(y,x)}{\pi_1(x)}, \frac{h(x,y)}{\pi_2(y)}\right)\pi_1(x)\pi_2(y)dxdy \\ &> a_1\frac{1}{2\pi_1(A_1)}\int_{A_1\times A_2} \pi_1(x)\pi_2(y)dxdy \\ &= \frac{1}{2}a_1\pi_2(A_2) \end{aligned} \tag{45}$$

where

$$a_1 = \inf_{x\in A_1, y\in A_2} \min\left(\frac{h(y,x)}{\pi_1(x)}, \frac{h(x,y)}{\pi_2(y)}\right). \tag{46}$$

Follow same argument to get for powered up distributions ξ_i's.

$$\begin{aligned} h'_{12} &= \frac{1}{2\xi_1(A_1)}\int_{A_1\times A_2} \min\left(1, \frac{\xi_2(y)}{\pi_1(x)}\frac{h(x)}{h(y)}\right)\xi_1(x)h(y)dxdy \\ &> \frac{1}{2}a_t\xi_2(A_2) \end{aligned} \tag{47}$$

where

$$a_t = \inf_{x\in A_1, y\in A_2} \min\left(\frac{h(y,x)}{\xi_1(x)}, \frac{h(x,y)}{\xi_2(y)}\right). \tag{48}$$

Note $\pi_2(A_2) = \xi_2(A_2)$, so the ratio h'_{12}/h_{12} is determined by a_t/a_1. However, for each $x \in A_1, y \in A_2$, $\frac{h(y,x)}{\xi_1(x)} = c\, t^d \frac{h(y,x)}{\pi_1(x)}$ due to Lemma 4.4 (c). So we have $a_t/a_1 = c\, t^d$ and hence

$$h'_{12} = c\, h_{12}\, t^d. \tag{49}$$

For an $m \times m$ stochastic matrix $P_c = (h_{ij})$, the spectral gap can be bounded from below (Peña, 2005) by

$$\mathbf{Gap}(P_c) \geq m \min_{i\neq j} h_{ij}. \tag{50}$$

Combine Equations (49) and (50) to finish the proof. □

5. Applications. There are three key parameters needed to be specified in applications, namely, the proportional of the long-range proposals s, the number of chains $(H+1)$ and the temperature ratio τ. The theoretically best value of s is $1/3$ because it maximizes the lower bound of the spectral gap of a Small-World chain (Guan and Krone, 2007). Indeed, in the following two examples, we use $s = 0.33$. Of course for a specific application one may tune s based on acceptance ratio. We note, however, s should keep constant during a MCMC run. To specify H and τ we suggest the following procedure: First, sample $\pi_t(x)$ and tune t until acceptance ratio of long-range proposal is high, say, larger than 0.2. Next, use Algorithm 2.2 to sample $\pi_t(x), \pi_{t/\tau}(x)$, tuning τ so that the acceptance ratio for long-range proposal of the sampling chain is > 0.2. H can be estimated by $[\log t/\log \tau]$. One may find burn-in and thinning helpful.

Lastly, the choice of long-range proposal is often problem-dependent. If the state space is $\mathbb{R}^d$, we recommend a heavy tailed distribution like Cauchy. When the state space is trees or graphs, a long-range proposal is hard to define – we suggest to compound local proposals of randomly many times to obtain a long-range proposal.

5.1. *Sampling Needles.* In this example, our target distribution is a mixture of normals, $f(x) = 0.5N(x;\mu_1,\Sigma_1) + 0.5N(x;\mu_2,\Sigma_2)$ where $\mu_1 = (0,0)^t$, $\mu_2 = (5,5)^t$ and $\Sigma_1 = \Sigma_2 = \left(\begin{smallmatrix} 0.01 & 0 \\ 0 & 0.01 \end{smallmatrix}\right)$. To see the minorization and drift condition hold for this example, see Atchadé (2010) and references therein.

Our local proposal is two dimensional ball with radii 0.1. The long-range proposal is two dimensional Cauchy with scale parameter 1. Local chains are trapped in either mode within $1,000,000$ iterations (data not shown). We use 6 chains with $\tau = 6$ with 1000 burn-in steps in each chain. The sampling chain ran $10,000$ steps, which makes the total iterations of the all 6 chains $81,000$ steps. Define region $A = \{x : x_1^2 + x_2^2 < 0.05^2\}$ and $p = Pr(X_n^{(6)} \in A)$, the probability that the sampling chain visits A. Figure 5.1 is the sample trace of a typical run, where after thinning of every 10 steps, the last $1,000$ samples of each chain were plotted.

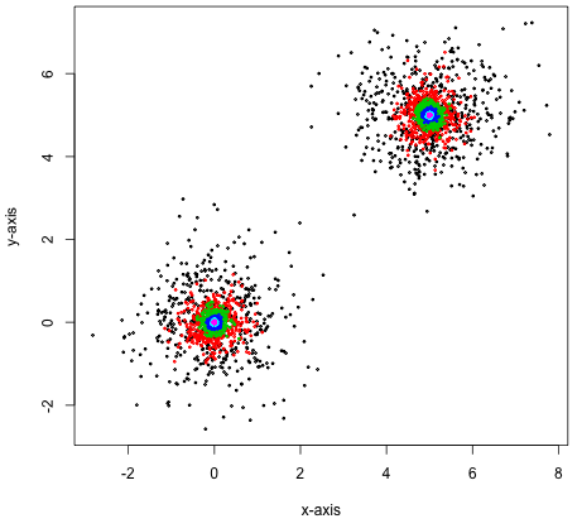

FIG 1. *Each color corresponds to different temperature* 6^k, $k = 5, 4, \ldots, 1, 0$. *The proportion of the sampling chain that visit region* A *is* 44.6%

We repeat above run 100 times and obtain statistics regarding $\hat{p}$. The mean is 0.50 and median is 0.49, the standard deviation is 0.08, and the 5 and 95 percentiles are 0.37 and 0.62 respectively. We repeat simulations for $\mu_2 = (25, 25)^t$ without modifying $l(x, \cdot)$ and $h(x, \cdot)$. With an extra chain and twice number of iterations, similar results are obtained (data not shown).

5.2. *Sampling Phylogenetic Trees.* Markov chain Monte Carlo algorithms plays an important role in (Bayesian) phylogenetic inference, perhaps through the wide-spread usage of software packages like MrBayes (Ronquist and Huelsenbeck, 2003) and PAML (Yang, 1997). Mossel and Vigoda (2005) argued that phylogenetic MCMC algorithms are misleading when data (nucleotides sequences) are generated by mixture of phylogenetic trees. Fixing branch lengths, they generated sequence data using two trees that are far apart (that is, local proposals can not reach from one to another in one step). They first showed that there is a valley in between the two trees used to simulated sequence data, and the valley becomes steeper when the sequence length (N) increases. Then they argued that existing local samplers takes exponentially long iterations (in N) to move from one mode to another. Their theoretical results is essentially the first part of the Theorem 3.1 in Guan and Krone (2007). In light of Guan and Krone (2007) and theories presented in this paper, we see that the slow mixing problem presented in Mossel and Vigoda (2005) can be simply resolved by a Small-World sampler, and better mixing can be achieved by a STEEP sampler.

In our simulation, we fix the branch lengths the same as those in (Mossel and Vigoda, 2005) with

inner branch lengths equal 0.1 and tip branch lengths equal 0.01. We use Jukes-Canter as the evolutionary model to compute likelihood of different tree topologies. Our local proposal is the nearest neighbor interchange (NNI) (c.f. Felsenstein, 2004), and the long-range proposals are simulated by compounding multiple (but random many) nearest neighbor interchanges. In this example, the minorization and drift condition hold because the state space is finite.

The five taxa example presented in (Mossel and Vigoda, 2005) is too simple for a simulation study. We simulate DNA sequence data (Rambaut and Grass, 1997) based on two generating trees (in Newick format) $A = (((((1,2),3),4),5),6),(7,8))$ and $B = (((((1,7),3),4),5),6),(2,8))$. Note we switch the position of taxa 2 and 7 and it takes 5 NNI to move from tree A to tree B. We found it difficult to simulate sequence data from the two trees that results in similar likelihood on both, so we simulate sequences of length 1000 from tree A and switch the sequences 2 and 7 to obtain new sequences and concatenate two sets of sequences together. By doing so, we ensure that tree A and B have the same likelihood.

We first ran a simple Metropolis-Hastings algorithm and plot the distance between taxa 1 and 2, the local chain were trapped within one mode during $1,000,000$ steps (data not shown). We then ran STEEP sampler of 4 chains with $\tau = 10$, each chain ran $50,000$ steps with $5,000$ steps of burn-in. After thinning every 10 steps, the last 5000 were plotted for each chain. The left panel shows the likelihood trace plot of each chain. The right panel shows the distance (between taxa 1 and 2) trace plot. Clearly, the chain moves frequently between trees A and B.

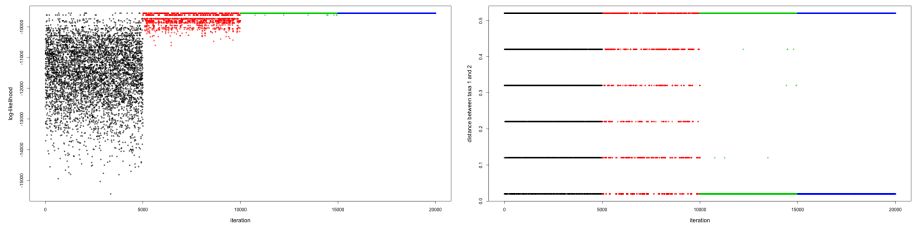

FIG 2. *Each color band corresponds to a different chain of temperature* 10^k, $k = 3, 2, 1, 0$. *The best log-likelihood of trees that is one local proposal (NNI) away from the generating trees is* 60.44 *smaller than those of generating trees. Only two generating trees appear in the last chain and the proportion of the two generating trees are* 44.0% *and* 56.0% *respectively.*

This toy example demonstrates that the STEEP performs well in mixture of trees where a local chain fails. Note in the example we fixed the branch lengths and evolutionary parameters. We invite authors of MrBayes, PAML and others to further investigate its performance when taking into account of branch lengths and evolutionary models.

6. Discussion. We have presented a new sampling algorithm, proved its ergodicity, and demonstrated its usefulness. The analysis of the spectral gap appears to be new. Although the theory is presented in the Euclidian state space, we believe it applies to more general state space as well. The STEEP algorithm bears similarities with the Equi-Energy sampler. One key difference is that

STEEP emphasizes the long-range proposal, through which the STEEP provides a new perspective on the advantage of using empirical distribution of tempering. We note that, at least in principle, the ergodic theorem and the analysis of the spectral gap apply to the Equi-Energy sampler. The STEEP also bears similarities with pure tempering methods such as MCMCMC. We point out three key differences. First, STEEP takes advantage of the empirical distributions, while the "swap" in tempering methods always use the current state of each chain. Second, the interaction of the different chains in STEEP is one-way – from high temperature chains to the lower ones. Since the higher temperature chains are not affected by the lower temperature chain, the exploring is more efficient. Third, tempering relies on local proposals to be more efficient on a *flattened* distribution, while STEEP relies on long-range proposals to be more efficient on a *fattened* distribution.

The "Powering-up" is convenient to obtain flattened (or fattened) alternative distributions. However, for certain type of models, e.g., Ising model and its generalization, Potts model, powering-up could run into problems because there might exists a phase transition at a critical temperature (Bhatnagar and Randall, 2004). When that happens, distributions of above and below the critical temperature may have little similarity; thus it becomes a moot point to use empirical distribution of one temperature to generate long-range proposals for another. To circumvent the "phase transition" one should discard powering-up scheme. Instead, one may use the "multi-set sampler" (Leman *et al.*, 2009) to *smooth* a distribution to achieve a similar effect as tempering. A multi-set sampler augments the state space from $\mathbb{R}^d$ to $\mathbb{R}^{d\times m}$ so that the current state is a vector $(x_1, \cdots x_m)$. Define $\pi'(x_1, \cdots, x_m) = \frac{1}{m}(\pi(x_1) + \cdots + \pi(x_m))$. This averaging effectively gives a smoother marginal distribution $\pi'_1(\cdot)$ compared to $\pi(\cdot)$. And we can control the smoothness by varying m.

In this paper, we focus on the mixing between different modes because within each mode local proposals guarantee rapidly mixing. However, when there exists a mode that is highly correlated among certain dimensions, one needs to fine tune the local proposals. Incorporating certain adaptive sampling scheme such as Craiu *et al.* (2009) for local proposal into STEEP might be desirable.

REFERENCES

Andrieu, C., Jasra, A., Doucet, A., and Del Moral, P. (2007). On nonlinear markov chain monte carlo via self-interacting approximations. *Technical Report*.

Andrieu, C., Jasra, A., Doucet, A., and Del Moral, P. (2008). A note on convergence of the equi-energy sampler. *Stochastic Analysis and Applications*, **26**, 298–312.

Atchadé, Y. (2010). A cautionary tale on the efficiency of some adaptive monte carlo schemes. *Ann. Appl. Probab.*, **20**, 116154.

Atchadé, Y. F. and Liu, S. J. (2006). Discussion of equi-energy sampler. *Annals of Statistics*, **34**(4), 1620–1628.

Atchadé, Y. F. and Rosenthal, J. S. (2005). On adaptive markov chain monte carlo algorithms. *Bernoulli*, **11**, 815–828.

Bhatnagar, N. and Randall, D. (2004). Torpid mixing of simulated tempering on the Potts model. In *Proceedings of the 15th Annual ACM-SIAM Symposium on Discrete Algorithms (New Orleans, LA)*, pages 478–487.

Brockwell, A., Del Moral, P., and Doucet, A. (2010). Sequentially interacting markov chain monte carlo. *Ann. Statist.*, **38**, 3387–3411.

Craiu, R. V., Rosenthal, J. S., and Yang, C. (2009). Learn from thy neighbor: Parallel-chain adaptive mcmc. *Preprint*.

Davidson, J. and de Jong, R. (1997). Strong laws of large numbers for dependent heterogeneous processes: a synthesis of recent and new results. *Econometric Rev.*, **16**, 251–279.

Del Moral, P. and Doucet, A. (2010). Interacting markov chain monte carlo methods for solving nonlinear measure-valued equations. *Ann. Appl. Probab.*, **20**, 593639.

Felsenstein, J. (2004). *Inferring Phylogenies*. Sinauer Associates, Sunderland, Mass.

Fort, G., Moulines, E., and Priouret, P. (2010). Convergence of adaptive mcmc algorithms: ergodicity and law of large numbers. *Technical Report, Paris Tech.*

Geyer, C. J. (1991). Markov chain Monte Carlo maximum likelihood. In E. M. Keramides, editor, *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, pages 156–163. Interface Foundation, Fairfax Station.

Guan, Y. and Krone, S. M. (2007). Small-world mcmc and convergence to multi-modal distributions: From slow mixing to fast mixing. *Annals of Applied Probability*, **17**, 284–304.

Guan, Y., Fleißner, R., Joyce, P., and Krone, S. M. (2006). Markov chain Monte Carlo in small worlds. *Stat. Comput.*, **16**, 193–202.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika*, **57**, 97–109.

Jarner, S. F. and Roberts, G. O. (2007). Convergence of heavy-tailed monte carlo markov chain algorithms. *Scandinavian Journal of Statistics*, **34**, 781815.

Kannan, R., Lovász, L., and Simonovits, M. (1995). Isoperimetric problems for convex bodies and a localization lemma. *Discrete Comput. Geom.*, **13**, 541–559.

Kou, S. C., Zhou, Q., and Wong, W. H. (2006). Equi-energy sampler with applications in statistical inference and statistical mechanics, with discussion. *Annals of Statistics*, **34**, 1581–1619.

Lawler, G. F. and Sokal, A. D. (1988). Bounds on the L^2 spectrum for Markov chains and Markov processes: a generalization of Cheeger's inequality. *Trans. Amer. Math. Soc.*, **309**, 557–580.

Leman, S. C., Chen, Y., and Lavine, M. (2009). The multiset sampler. *Journal of the American Statistical Association*, **104**(487), 1029–1041.

Lovász, L. and Simonovits, M. (1993). Random walks in a convex body and an improved volume algorithm. *Random Structures Algorithms*, **4**, 359–412.

Lovász, L. and Vempala, S. (2003). The geometry of logconcave functions and an $O^*(n^3)$ sampling algorithm. Microsoft Research Tech. Rep. MSR-TR-2003–4. Available at : http://www-math.mit.edu/~vempala/papers/logcon-ball.ps.

Madras, N. and Randall, D. (2002). Markov chain decomposition for convergence rate analysis. *Ann. Appl. Probab.*, **12**, 581–606.

Madras, N. and Zheng, Z. (2003). On the swapping algorithm. *Random Structures and Algorithms*, **1**, 6697.

Marinari, E. and Parisi, G. (1992). Simulated tempering: a new Monte Carlo scheme. *Europhysics Letters*, **19**, 451–458.

Mathé, P. and Novak, E. (2007). Simple monte carlo and the metropolis algorithm. *J. Complex.*, **23**(4-6), 673–696.

Metropolis, N., Rosenbluth, A. E., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *J. Chem. Phys.*, **21**, 1087–1091.

Meyn, S. P. and Tweedie, R. L. (1993). *Markov Chains and Stochastic Stability*. Springer-Verlag.

Mossel, E. and Vigoda, E. (2005). Phylogenetic MCMC Algorithms Are Misleading on Mixtures of Trees. *Science*, **309**(5744), 2207–2209.

Neal, R. M. (2001). Annealed importance sampling. *Statistics and Computing*, **11**(2), 125–139.

Peña, J. M. (2005). Exclusion and inclusion intervals for the real eigenvalues of positive matrices. *SIAM Journal on Matrix Analysis and Applications*, **26(4)**, 908–917.

Predescu, C., Predescu, M., and Ciobanu, C. V. (2004). The incomplete beta function law for parallel tempering sampling of classical canonical systems. *J.CHEM.PHYS.*, **120**, 4119.

Rambaut, A. and Grass, N. C. (1997). Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput. Appl. Biosci.*, **13**(3), 235–238.

Roberts, G. O. and Rosenthal, J. S. (1997). Geometric ergodicity and hybrid Markov chains. *Electron. Comm. Probab.*, **2(2)**, 13–25.

Roberts, G. O. and Rosenthal, J. S. (2004). General state space markov chains and mcmc algorithms. *Probability Surveys*, **1**, 20–71.

Roberts, G. O. and Tweedie, R. L. (2001). Geometric L^2 and L^1 convergence are equivalent for reversible Markov chains. *J. Appl. Probab.*, **38A**, 37–41.

Ronquist, F. and Huelsenbeck, J. P. (2003). Mrbayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, **19**, 1572–1574.

Royden, H. L. (1988). *Real Analysis, second edition*. Collier-Macmillan, London.

Rudolf, D. (2009). Explicit error bounds for lazy reversible markov chain monte carlo. *Journal of Complexity*, **25**(1), 11 – 24.

Woodard, D. B., Schmidler, S. C., and Huber, M. (2008). Conditions for rapid mixing of parallel and simulated tempering on multimodal. *Annals of Applied Probability*.

Yang, Z. (1997). Paml: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.*, **15**, 555–556.

BAYLOR COLLEGE OF MEDICINE
CHILDREN'S NUTRITION RESEARCH CENTER
1100 BATES ROOM 2070
HOUSTON, TX 77030
E-MAIL: yongtaog@bcm.edu

UNIVERSITY OF CHICAGO
ECKHART HALL ROOM 126
5734 S. UNIVERSITY AVENUE
CHICAGO, IL 60637
E-MAIL: mstephens@uchicago.edu